DOCUMENT RESUME

ED 423 307 TM 029 115

AUTHOR Schumacker, Randall E.
TITLE Comparing Measurement Theories.
PUB DATE 1998-04-15
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Computer Software; *Error of Measurement; *Generalizability Theory; Measurement Techniques; *Raw Scores; *Test Theory
IDENTIFIERS *Rasch Model

ABSTRACT

In comparing measurement theories, it is evident that the awareness of the concept of measurement error during the time of Galileo has lead to the formulation of observed scores comprising a true score and error (classical theory), universe score and various random error components (generalizability theory), or individual latent ability and error estimates (latent trait theory). The definition of a true score and the definition of measurement error separates measurement theories. Students who need practical applications can progress through the traditional software for each of these theories. Using the software for the various approaches will give students an understanding of the measurement theories, scaling, objective measurement, and the different definitions of true score and error. (Contains 2 tables and 21 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 423 307

TM

COMPARING MEASUREMENT THEORIES

RANDALL E. SCHUMACKER
UNIVERSITY OF NORTH TEXAS

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Randall Schumacker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper Presented at the American Educational Research Association
April 15th, 1998
Symposium
Institute for Objective Measurement Educators Forum

TM029115

COMPARING MEASUREMENT THEORIES

Randall E. Schumacker
University of North Texas

Traub (1997) highlighted several major concepts in classical test theory: correction for attenuation, Spearman-Brown Prophecy formulas (KR20/KR21), and Guttman's lower bounds to reliability. Absent in his presentation is the initial understanding and awareness that sparked the historical trends in measurement, namely, the acknowledgment of error in measurement back in the days of Astronomer Galileo. Early in this history of measurement, the awareness that an observed score had as components a true score plus error launched many of the other historical developments in measurement.

Pearson's correlation coefficient formula (1896) introduced a way to inter-correlate items that could measure a construct or variable of interest. Spearman (1904) used the inter-item correlation matrix to develop factor analytic methods. The notation that factor scores represented true scores and loaded on a common factor was second only to examining the residual factor that contained potentially other specific factors and error. Hence the notation $X = T + E$. Factor analytic methods were specifically developed to reproduce the original correlation in contrast to principal components which maximized variance (Jöreskog, 1979). Eventually, the concept of error was examined in relation to its' presence or absence in correlation, variance-covariance, and regression analyses (Werts, Rock, Linn, & Jöreskog, 1976). Jöreskog (1973) extended this notion of error in observed scores to create a statistical program which incorporates measurement error into the statistical analysis of data.

Classical and Generalizability Theory

In classical theory, corrections for attenuation in correlation coefficient and separate reliability coefficients for different testing conditions were formulated. Each reliability coefficient was sample based and yielded a single standard error of measurement which applied to all scores. Cronbach's early work reflected this formulation of reliability (1947; 1951). Later, Cronbach, Gleser, Nanda, & Rajaratnam (1972), realized that separate reliability coefficients conceptualizing error in measurement actually reflected the generalizability of scores given testing design conditions. Cronbach and his colleagues conceived the idea of an analysis of random effects variance components, in contrast to the fixed effects ANOVA. Hence the idea of a well defined domain or universe, the random sampling of items, and resultant generalizability of scores. Many researchers have incorrectly described G-theory as an analysis of variance procedure when in fact it is rooted in random effects variance components grounded in the factorial design work of Fisher (1925) and expanded in the 1940's by Hoyt (1941). Closely linked to this realization is "Expected Mean Squares" resulting from the Cornfield and Tukey algorithm (Cornfield & Tukey, 1956) to estimate expected variance components. The step-by-step procedure for the Cornfield-Tukey algorithm is illustrated in Dayton (1970). Overall, G-theory partitions random effects variance components such that all of the individual measurement errors could be modeled at the same time, e.g., internal consistency, test-retest, parallel forms. Consequently, multiple analyses of facets can lead to an alternative definition of error (Lindquist, 1953). Given this perspective, G-theorists define the conditions of measurement, crossed or nested, then seek to determine which conditions yield dependable scores in a sample of persons

(G-study vs. D-study). The dependability of scores is based upon a decision pertaining to the number of testing conditions needed. For example, number of randomly parallel forms of a test, number of testing occasions, number of raters, or number of items sampled. Each of these testing conditions yield random effects variance components with error variance.

Latent Trait Theory

Bock (1997) articulated that Thurstone's (1925) early work reflected principles based on item response theory as conceptualized today. Thurstone formulated a score model expressed as the probability of success on a given item as a function of a continuous variable, i.e., ability, expressed as an absolute score scale. His goal was to develop an objective scale. In modern IRT approaches, the continuum variable is defined as a latent variable or ability impacting the probability of a correct response to an item. Two important analytical features emerged from this, namely, (1) each item can be calibrated on a scale with a unique error component and (2) each person can be calibrated on a scale with a unique error component. No longer was a group based single error of measurement (SEM) applied to all examinee scores, rather each individual and item had a unique error term. This was called sample free and item free estimation. Hence we have the "logit" unit of measurement, which is related to Fisher's original "z" transformation by 2p - 1 (Fisher & Yates, 1938), but derived from the Newton-Raphson maximum liklihood iterative method (Fisher, 1925). A logit to proportions comparison table can be found in Wright and Stone (1979, p.36).

Later developments by Lord (1952) and Birnbaum (1957), namely : logistic item response models replacing normal ogive models, MLE estimation using all information in the item response pattern, formulation of an item information function, and the introduction of a response model to include guessing, further served to enhance our understanding of the relationship between observed score, latent ability, and error of measurement.

Georg Rasch (1961) provided yet another perspective on latent ability estimation and objective measurement. The Rasch model postulated that item difficulty and person ability calibration alone was consistent, sufficient, and efficient. Rasch model parameters are estimated consistently using MLE in the conditional distribution of the item responses. Realization that each individual has their own logit ability estimate (latent ability) and error provided objective measurement scaling.

Summary

In comparing the measurement theories, one quickly realizes that the awareness of the concept of measurement error during Galileo's time has lead to the formulation of observed scores comprising a true score and error (classical theory), universe score and various random error components (generalizability theory), or individual latent ability and error estimates (latent trait theory). The definition of a true score and the definition of measurement error uniquely separates our understanding of the measurement theories, as does the assumptions of each theory (see Tables 1 & 2).

Students requiring practical applications can progress through the traditional software which yields a number correct (or percent correct), item difficulty, discrimination, standard error of measurement, and the individual calculations for reliability using ITEMAN or SPSS. Scores

are interpreted using the group based SEM value. These and other practical measurement topics are introduced using the Instructional Topics in Educational Measurement Series (NCME, 1997).

Progression to G-theory using GENOVA or SAS permits the practical understanding of how item sampling, occasions of testing, alternate forms of a test, and rating designs can be analyzed at the same time to determine which set of testing conditions would yield dependable scores. The concept of a universe score and these multiple random effects variance components as sources of measurement error are now better understood by students.

Problems with the measurement of individual performance can now be discussed and articulated (sample dependency, ordinal nature of test scores, absence of a continuous equal interval scale) to further the students understanding of the need for objective measurements. Software programs such as Rascal (Rasch model) or Ascal (IRT model) can easily be used after experience with ITEMAN (classical model) because the program format is similar. Experience using BIGSTEPS or BIGSCALE provides additional understanding of the Rasch model and diagnosing misfit, while experience using Bilog or Multilog provides an understanding of the IRT 1pl, 2pl, and 3pl models, which include item difficulty, item discrimination, and item guessing parameters, respectively. At this point, students should better understand the unique ability and error estimates derived in latent trait theory, as well as, differences in Rasch and IRT models. Other scoring and scaling methods can also be discussed (Likert, partial credit, graded response) and associated software examples presented using BIGSCALE, BIGSTEPS, or FACETS. Students have now gained an understanding of the measurement theories, scaling, objective measurement, and especially the different definitions of true score and error.

REFERENCES

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision making problems* (Series Report Number 58-16, Project Number 775-23). Randolph Air Force Base, Texas: United States Air Force School of Aviation Medicine.

Bock, D.R. (1997). A Brief History of Item Response Theory. *Educational Measurement: Issues and Practices, 16*(4), 21-33.

Cornfield, J. & Tukey, J.W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics, 27*, 907-949.

Cronbach, L.J. (1947). Test reliability: Its meaning and determination. *Psychometrika, 12*(1), 1-16.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 292-334.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley & Sons.

Dayton, C.M. (1970). *The Design of Educational Experiments*. New York, NY: Mcgraw-Hill.

Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society, 22*, 699-725.

Fisher, R.A. & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. New York: Hafner.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153-160.

Jöreskog, K.G. (1973). *A general method for estimating a linear structural equation system*. In A.S. Goldberger and O.D. Duncan (Eds): *Structural Equation Models in the Social Sciences*. New York: Seminar Press, 85-112.

Jöreskog, K.G. (1979). *Chapter 1: Basic Ideas of Factor and Component Analysis*. In Magidson, J. (Ed.) *Advances in Factor Analysis and Structural Equation Models*. Cambridge, MA: Abt Books.

Lindquist, E.F. (1953). *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin.

Lord, F.M. (1952). A theory of test scores. Psychometric Monograph, 7.

National Council on Measurement in Education (1997). *Instructional Topics in Educational Measurement Series, 1987-1997.* NCME Publication Sales, 1230 17th St. N.W., Washington, DC 20036-3078.

Pearson, K. (1896). Mathematical contributions to theory of evolution. Part 3. Regression, heredity and panmixia. *Philosophical Transactions, A, 187,* 253-318.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4,* 321-324.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16,* 433-451.

Traub, R.E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practices, 16*(4), 8-14.

Werts, C.E., Rock, D.A., Linn, R.L., & Jöreskog, K.G. (1976). Comparison of correlations, variances, covariances, and regression weights with or without measurement error. *Psychological Bulletin, 83*(6), 1007-1013.

Wright, B.D. & Stone, M.H. (1979). *Best Test Design.* MESA Press: Chicago, Illinois.

Table 1. Comparison of Measurement Theory True Score and Error

Classical Theory	Observed Score = True Score + Error Sample Dependent Measures Single Group Based Error Term
Generalizability Theory	Observed Score = Universe Score + Multiple Sources of Error Sample Dependent Measures Partition Sources of Error Variance for Phi- and G-Coefficients
Latent Trait Theory	Rasch Model: logit +/- residual Where: logit = B - D or Ability minus Item Difficulty IRT Model: θ_g +/- error Where: θ_g = different ability estimates based on difficulty, discrimination, guessing, or distractor item parameters in model. Sample and Item Free Measures Individual ability and error calibrations

Table 2. Comparison of Measurement Theory Assumptions

Classical Theory	Error of Measurement = Discrepancy between examinee observed score and true score Score Interpretation: X +/- (SEM) Assumptions: a. Mean of the error scores in a population of examinees is zero b. Correlation between error and true scores in a population of examinees is zero. c. Correlation between error scores from two independent distributions or two testing occasions using the same test is zero.
Generalizability Theory	Error of Measurement = Different error variances depending on testing conditions Score Interpretation: S +/- (SEM) Where: S or the universe score depends on the testing conditions(facets). Assumptions: a. Measurement conditions (facets) reflect universe of generalizations b. Fixed and Random Facets; Crossed and Nested Designs determine different universes of admissible observations c. Random Effects permit generalization to universe while Fixed Effects permit generalization to only those conditions specified

Latent Trait Theory	Error of Measurement = Difference between observed and predicted response, i.e. residual Score Interpretation: Rasch: logit +/- (residual) IRT: θ +/- (error) Where: Score indicates probability of responding correctly to an item given latent model Assumptions: a. A latent trait (ability) accounts for dependence among items. b. Unidimensionality (dependence among items or number of latent traits needed to achieve local independence) c. Local Independence (independence among items at ability levels) d. Test-Free Measurement e. Sample-Free Measurement

Table 2 - continued.